Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W645–W648***
doi:10.1093/nar/gkm333

The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods

Sebastien Moretti[1], Fabrice Armougom[3], Iain M. Wallace[2], Desmond G. Higgins[2], Cornelius V. Jongeneel[1] and Cedric Notredame[3,*]

[1]Swiss Institute of Bioinformatics, Bâtiment Génopode, UNIL, CH-101 Lausanne, [2]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland and [3]Information Génomique et Structurale, CNRS UPR2589, Institute for Structural Biology and Microbiology (IBSM), Parc Scientifique de Luminy, 163 Avenue de Luminy, FR- 13288, Marseille cedex 09, France

Received January 31, 2007; Revised April 15, 2007; Accepted April 18, 2007

ABSTRACT

The M-Coffee server is a web server that makes it possible to compute multiple sequence alignments (MSAs) by running several MSA methods and combining their output into one single model. This allows the user to simultaneously run all his methods of choice without having to arbitrarily choose one of them. The MSA is delivered along with a local estimation of its consistency with the individual MSAs it was derived from. The computation of the consensus multiple alignment is carried out using a special mode of the T-Coffee package [Notredame, Higgins and Heringa (T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*. 2000; 302: 205–217); Wallace, O'Sullivan, Higgins and Notredame (M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res*. 2006; 34: 1692–1699)] Given a set of sequences (DNA or proteins) in FASTA format, M-Coffee delivers a multiple alignment in the most common formats. M-Coffee is a freeware open source package distributed under a GPL license and it is available either as a standalone package or as a web service from www.tcoffee.org.

INTRODUCTION

The computation of an accurate multiple sequence alignment (MSA) is central to a large number of bioinformatics analyses, ranging from phylogeny, profile construction, structure prediction and more recently sequence/structure activity relationship. Despite its importance, the MSA problem has not yet met with a definitive answer and a wide variety of alternative methods are currently available (3,4). All these methods are meant to address the same problem in different ways. In recent years, many efforts have been undertaken to characterize their relative accuracy but the overall outcome suggests that there is no such thing as a perfect MSA method, with each individual method having specific strengths and weaknesses. In practice, evaluation is made using structure-based MSAs as a standard of truth and the expected accuracy of a method is deduced from its ability to produce a structurally correct sequence alignment while using sequence information only. At least five such collections of reference alignments (5–8) have been established, and although some methods give better average results than others, one cannot know in advance which method will outperform the others on a given dataset. As such, it is always possible for the worst method to outperform all the others on a specific dataset. For the biologist, this makes it impossible to use anything other than a weak statistical argument (i.e. best method on average) to choose one method among the others when computing an alignment.

The design of meta-methods (or jury-based methods) is one way of addressing such situations in biology. Meta-methods are meant to combine the output of several alternative methods into one final output. They are based on the empirical reasoning that errors produced by independent prediction systems should not be consistent, therefore suggesting agreement as an indication of correctness. Such an approach was successfully used in the field of gene predictions (9) or for secondary structure predictions (10). Combining alignments, however, is less simple than building consensus prediction and it is only in 1999 that an effective strategy was proposed by Bucka-Lassen (11). An alternative to the Bucka-Lassen strategy, using consistency, was later introduced in the T-Coffee (1) algorithm. Recently, this algorithm was further modified in order to address the problem of combining alternative MSAs into one (2). T-Coffee (1) is a progressive consistency-based algorithm that compiles an alignment

*To whom correspondence should be addressed. Tel: +33 491 106 486; Fax: +33 491 106 489; Email: cedric.notredame@europe.com

on the basis of its consistency with a collection of pairwise constraints. In practice, the constraints correspond to pairs of residues that could end up aligned in the final alignment. These constraints, however, are not necessarily all compatible with one another and the goal of the algorithm is to fit as many as possible within the final alignment, while discarding those that were hopefully biologically less relevant. The term consistency refers to the notion that one tries to compute the alignment having the highest possible consistency with the constraint list. This notion was introduced by Gotoh (12) and later re-used in several algorithms (13,14). In 2000, Notredame *et al*. (1) described a variation of the progressive algorithm using consistency as a scoring scheme. This combination proved quite successful and is now at the core of several MSA packages (15–17). In its default mode, T-Coffee uses, as a list of constraints, all the pair-wise matches extracted from a compilation of all possible global pairwise alignments and the 10 best local alignments from each pair of sequences. Yet, this is merely one of the possible recipes to assemble such a list of constraints, and alternatives are possible. For instance, ProbCons (16) uses suboptimal pairwise global alignments (as emitted by an HMM with posterior decoding); PCMA (15) uses pairwise profile comparisons and Expresso (18) uses a mixture of sequence and structure-based alignments. Following the same principle, it is also possible to generate alternative MSAs and compile them into a single list of constraints. This latest approach forms the basis of M-Coffee (2), where eight MSA methods are used to generate alternative MSAs. Extensive benchmarking showed that this combination results in a modest but consistent improvement over each individual method, with M-Coffee producing the best scoring alignment on two of three of the datasets contained in BaliBase (5), Prefab (6) and Homstrad (2).

Another interesting by-product of alignment combination is the possibility of estimating the local consistency between the final alignment and the individual alignments. This amounts to measuring, for every residue, the fraction of individual alignments that support its position in the final alignment. This measure is named the CORE index (Consistency of Overall Residue Evaluation) and was shown to be very informative with respect to the overall alignment accuracy (19). These initial reports recently gained further support thanks to some extensive analysis carried out by Sonhammer *et al*. (20) whose results indicate that the consistency between an MSA and a pre-computed collection of alternative alignments gives very reliable information with respect to the structural correctness of that alignment. As such, the local consistency measure appears to be one of the most reliable predictors of alignment accuracy available today.

The server we present here computes an alignment with eight of the most commonly used MSA packages. It then outputs a consensus alignment along with a CORE-based local evaluation that can either be color-coded or ASCII based. Two mirrors of these services currently run on separate clusters: one at the Swiss Institute of Bioinformatics on the Vital-IT framework, the other at the CNRS in Marseilles, France. Both mirrors can be accessed via the T-Coffee homepage: www.tcoffee.org and extra mirrors should be added in the close future.

METHODS

Primary library: computation of the initial MSAs

The principle of M-Coffee is to compute several alternative multiple alignments in order to combine them into one consensus alignment. By default, eight methods were chosen for this purpose: PCMA (Version 2.0) (15), POA (Version 2.0) (21), Dialign-t (Version 0.2.1) (22), MAFFT (Version 5.431, L-INS-i) (17), Muscle (Version 3.6), ProbCons (Version 1.2), ClustalW (23) and T-Coffee (1). Apart from MAFFT that is used in its most accurate mode (mafft- - localpair- -maxiterate 1000) all the methods are run on the initial dataset using the default parameters. This produces an MSA that is then turned into a T-Coffee primary library. All these libraries are then combined in order to generate an MSA.

M-Coffee alignment computation

In order to compute the final alignment, the server runs the following command:

t_coffee <seq> -method poa_msa, dialignt_msa, mafft_msa, clustalw_msa, muscle_msa, probcons_msa, t_coffee_msa, pcma_msa.

Using the M-Coffee server

The server can be accessed at www.tcoffee.org. Following the M-Coffee link will either take the user to the regular or advanced mode. The regular mode merely requires the user to cut and paste a set of sequences in FASTA format. The advanced mode (Figure 1) offers more possibilities and guides the user with a series of bulleted points:

(i) Cut and paste your sequences. Sequences should be in FASTA format. Duplicated names are now supported although not recommended.
(ii) Alignment computation. This section defines the way the primary library is computed. For instance, selecting only lalign_id_pair and slow_pair will lead to the computation of a regular T-Coffee MSA. The lower section (xxx_msa) displays the list of available MSA methods. Selecting only one of these methods will generate the corresponding alignment. Selecting several methods (or all of them, as in the regular mode displayed on Figure 1) will lead to a consensus T-Coffee MSA. If the MSA method one wants to combine is missing on this form, another server named 'Combine' should be used (accessible from www.tcoffee.org). The 'Combine' server works on the same principle as M-Coffee but does not compute the MSAs itself and requires the user to cut and paste pre-computed MSAs. At this point it should be used if one wants to incorporate specific constraints or structure-based sequence alignments.
(iii) Output. The Output section makes it possible to control the output format. The most notable

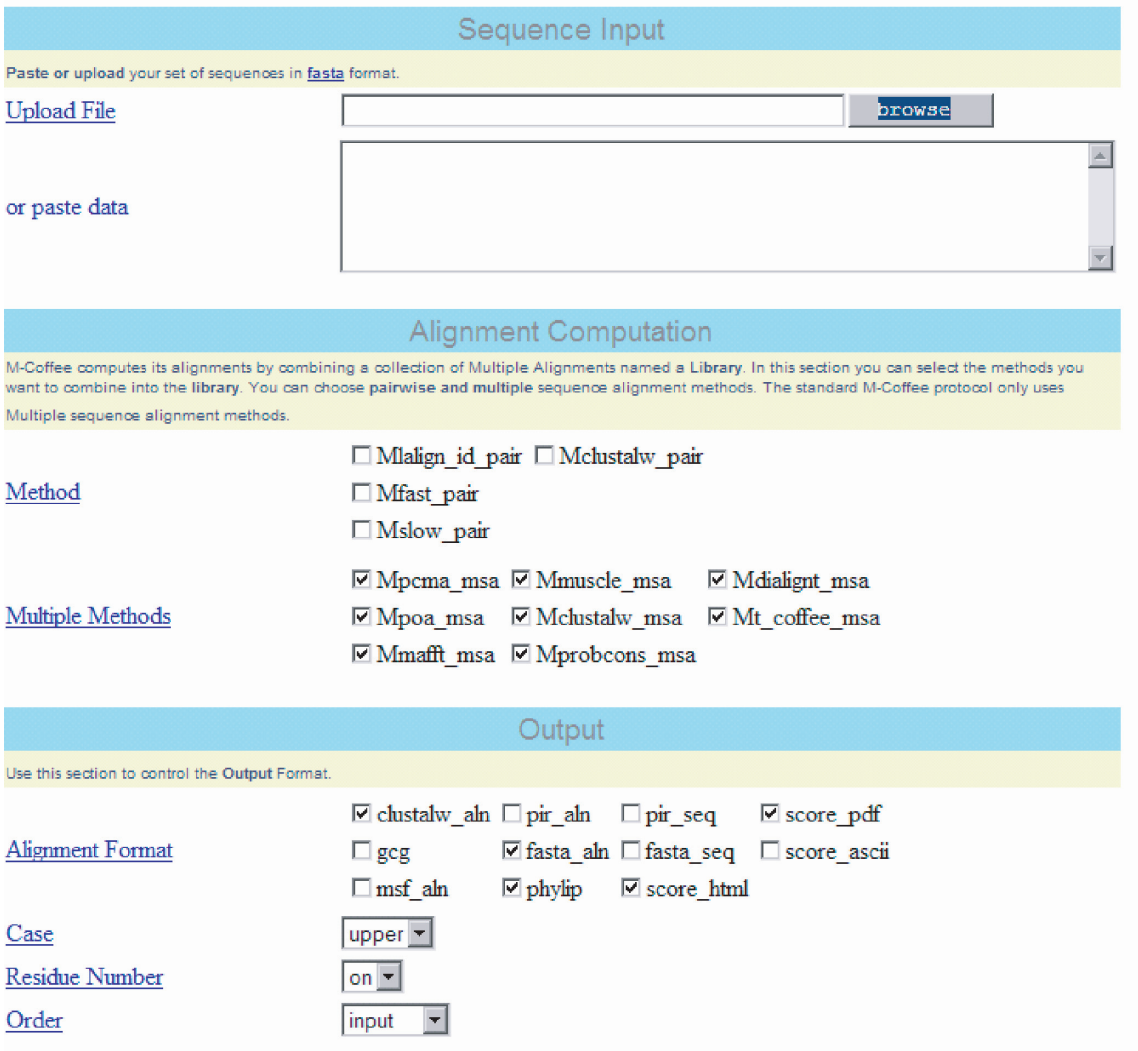

Figure 1. Method selection on the advanced M-Coffee server form. Each check box corresponds to either a pairwise (_pair) or a multiple sequence alignment method (_msa). Users should choose their methods of choice in order to combine them.

element is score_html that will cause the server to produce a colored version of the final alignment (Figure 2). In this output, residues are individually colored according to the consistency of their alignment with the T-Coffee library. Residues in red are in perfect agreement with every constituting multiple alignment while those in blue have the lowest agreement (i.e. the lowest support in the individual MSAs). Previous analysis indicates that 90% of the residues having a score of 7 or higher (dark yellow, orange and red) are correctly aligned (24). A text version of this output is available as score_ascii where each residue is replaced with its consistency estimation on a scale between 0 and 9 (9 corresponding to the red-brick residues in the color-output). These score_ascii files can be used to process multiple alignments (block extraction) using seq_reformat, one of the utilities distributed along with T-Coffee. For this purpose, users can download their alignment, the score_ascii file and use the

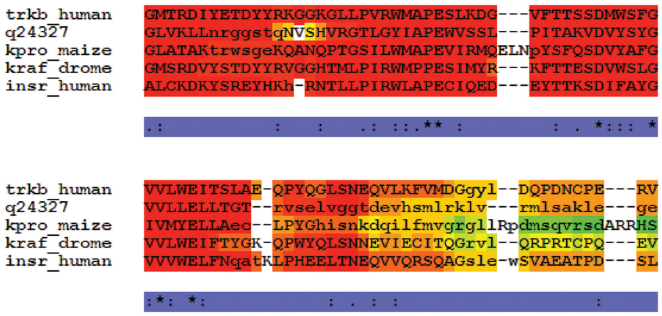

Figure 2. Typical colored output. This output was obtained by using the kinase1_ref5 from BaliBase. Correctly aligned residues (as judged from the reference) are in upper case, non-correct ones are in lower case. In this colored output, each residue has a color that indicates the agreement of the individual MSAs with respect to the alignment of that specific residue. Dark red indicates residues aligned in a similar fashion among all the individual MSAs; blue indicates a very low agreement. Dark yellow, orange and red residues can be considered to be reliably aligned.

command line version of T-Coffee with the following syntax:

t_coffee -other_pg seq_reformat -in <aln> -struc_in <score_ascii> -struc_in_f number_aln -action +keep '[5-9]'

Where <aln> is the name of the alignment and <score_ascii> the name of the score_asccii file. This syntax will replace by a gap ('-') every residue having an ascii_score lower than 5 (green and blue residues on the colored output).

CONCLUSION AND FUTURE DEVELOPMENTS

M-Coffee provides biologists with a useful alternative to the *a priori* choice of an MSA method. Although M-Coffee does not entirely solve the question of which method should be used, its local scoring scheme makes it easier to read the alignment and determine which portions are the most informative. Further developments will include making more methods available, as well as making it possible to combine sequences and structures, using the Expresso protocol.

ACKNOWLEDGEMENTS

The development of the server was supported by CNRS (Centre National de la Recherche Scientifique), by the Vital-IT frame work and by the European Union (ICGR-SIB FP6-026204). We thank Dr Pierre Pontarotti and Dr Vladimir Saudek for stimulating discussions. D.G.H. and I.M.W. are grateful to Science Foundation Ireland for funding. Funding to pay the Open Access publication charges for this article was provided by Département des Science de la Vie, Centre National de la Recherche Scientifique.

Conflict of interest statement. None declared.

REFERENCES

1. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
2. Wallace,I.M., O'Sullivan,O., Higgins,D.G. and Notredame,C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.
3. Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–373.
4. Wallace,I.M., Blackshields,G. and Higgins,D.G. (2005) Multiple sequence alignments. *Curr. Opin. Struct. Biol.*, **15**, 261–266.
5. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
6. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797. Print 2004.
7. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
8. Raghava,G.P., Searle,S.M., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.
9. Allen,J.E. and Salzberg,S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.
10. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. and Barton,G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
11. Bucka-Lassen,K., Caprani,O. and Hein,J. (1999) Combining many multiple alignments in one improved alignment. *Bioinformatics*, **15**, 122–130.
12. Gotoh,O. (1990) Consistency of optimal sequence alignments. *Bull. Math. Biol.*, **52**, 509–525.
13. Vingron,M. and Argos,P. (1991) Motif recognition and alignment for many sequences by comparison of dot-matrices. *J. Mol. Biol.*, **218**, 33–43.
14. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment [In Process Citation]. *Bioinformatics*, **15**, 211–218.
15. Pei,J., Sadreyev,R. and Grishin,N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.
16. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
17. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
18. Armougom,F., Moretti,S., Poirot,O., Audic,S., Dumas,P., Schaeli,B., Keduas,V. and Notredame,C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.
19. Notredame,C. and Abergel,C. (2003) In Andrade,M. (ed.), *Bioinformatics and Genomes: Current Perspectives*, Horizon Scientific Press, pp. 30–50.
20. Lassmann,T. and Sonnhammer,E.L. (2005) Automatic assessment of alignment quality. *Nucleic Acids Res.*, **33**, 7120–7128.
21. Lee,C., Grasso,C. and Sharlow,M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.
22. Subramanian,A.R., Weyer-Menkhoff,J., Kaufmann,M. and Morgenstern,B. (2005) DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.
23. Thompson,J., Higgins,D. and Gibson,T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4690.
24. Abergel,C., Coutard,B., Byrne,D., Chenivesse,S., Claude,J.B., Deregnaucourt,C., Fricaux,T., Gianesini-Boutreux,C., Jeudy,S. *et al.* (2003) Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. *J. Struct. Funct. Genomics*, **4**, 141–157.